WeRateDogs Data Wrangling Report

About Datasets

The dataset we'll be wrangling is the tweet archive of Twitter user @dog_rates (https:twitter.com/dog_rates), also know as WeRateDogs. WeRateDogs is a Twitter account that rates dogs and also provides a description using humorous comment about the dog.

Gathering Twitter archive file

I setup a Twitter's developer account by going through the portal. From there I generated the Consumer API keys, the Access Token and Access Token Secret.
I downloaded the tsv file using request library then imported the tweet image prediction tsv file into a DataFrame

```
response = requests.get(url)
with open('image_predictions.tsv', mode = 'wb') as file:
   file.write(response.content)

# Import the tweet image predictions TSV file into a DataFrame
img_df = pd.read_csv('image_predictions.tsv', sep='\t')
```

I also gathered WeRateDogs entire tweet data using Twitter API; and stored the JSON data in a file called tweet_json.txt. I created a DataFrame called status_df from this JSON

Data Assessment

For visual assessment, we opened the twitter_archive_enhanced.csv and image_predictions.tsv into Excel. As such we were able to scroll through them while looking for quality and tidiness issues. We were able to find two quality and 2 tidiness issues.

Quality

- Text column of twitter archive non truncated text as opposed to displayable text
- Html tags in source column of twitter archive as opposed to utility name. for example <a href=''http:….iPhone

Tidiness

- Twitter archive data illustrating retweets will have empty retweeted_staus_id, retweeted_staus_user_id and retweeted_status_timestamp columns. These columns can be dropped.

- Doggo, floofer, pupper and puppo columns should be merged into one column named "stage"

I used pandas "info" function on twitter_arch to find wrong datatypes and other quality issues. I also used "value counts" function to look up the range of the name column.
I used "query" method in order to see if there were tweets that had more than one dog-stage name mentioned. In the process I discovered a few more quality and tidiness issues.

Quality

- tweet_id of twitter_arch are missing in img_df table
- retweets which are duplicates
- wrong datatypes for in_reply_status_id, in_reply_to_user_id and timestamp columns
- some records have more than one dog stage name

Tidiness

- "breed" column should be added in twitter_arch
- Retweet_count and favorite_count columns from status_df table should be joined with twitter_arch table

Cleaning Data

I created a copy of twitter_arch table in order to perform the cleaning process. For each quality and tidiness issue I performed a programmatic data cleaning. I converted the datatypes of source and newly created stage columns of archive_clean to category datatype.

Storing Data

After performing all necessary cleaning process, I stored the _archive_clean DataFrame into "twitter_archive_master.csv" file